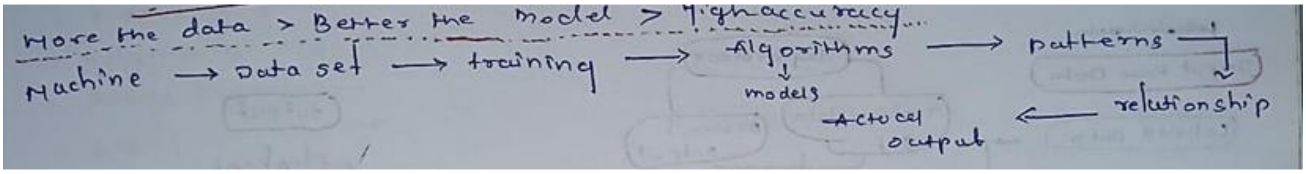
UNIT 4 – Introduction to Machine Learning in Bioinformatics

Machine Learning –

- A branch of artificial intelligence (AI) and computer science
- focuses on the use of data and algorithms to imitate the way that humans learn gradually, improving its accuracy
- Steps of machine learning –
  1. Identify and create the appropriate data set
  2. Perform computation to learn
  3. Required rules pattern and relations



- How does machine learning work ?

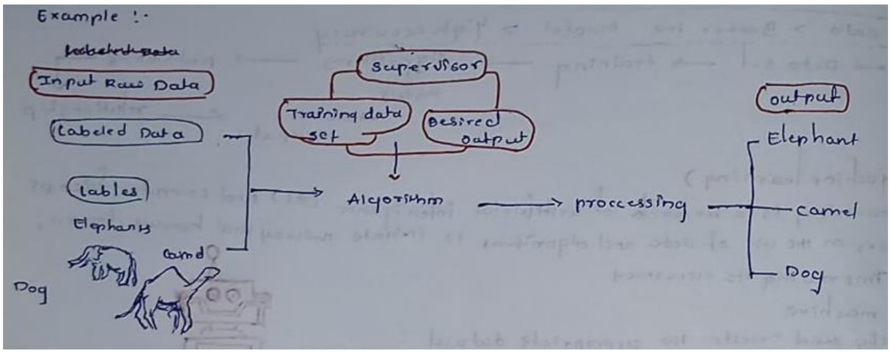
Sr. No.	Steps	Process	Description
1	A decision process	ML Algorithms used to make up prediction or classification	Based on input data (labelled / unlabelled), algorithm produces an estimate about a pattern in the data.
2	An Error Function	Evaluates the prediction of the model	If there are known examples ↓ Error function can make a comparison to assess the accuracy of the model.
3	A Model Optimization Process	Iterative 'evaluate and optimize' process	If model fits better to the data points in the training set ↓ Weights adjusted to reduce the discrepancy between known example and model estimate. ↓ Updating weights autonomously until a threshold of accuracy has been met.

- Types of machine learning –
  1. Supervised ML
  2. Unsupervised ML
  3. Semi-supervised ML
  4. Reinforcement learning

1. Supervised Machine Learning –

- **Def –** When a model gets trained on a 'labelled data set' – which have both input and output parameters.
- **Training dataset = Samples + Labels**
- Algorithms learn to map points between inputs and correct outputs
- It has both training and validation data sets labelled.
- **Example – Image classifier –**

Datasets of labelled images of cats and dogs fed to the algorithm → Machine learns to differentiate and classify → New Input: new dog / cat image, machine will use learned algorithms and predict whether it is a cat or dog



2 main categories of Supervised learning

Classification		Regression
Predicting categorical target variables was represent discrete classes or labels	**Deals with**	Predicting continuous target variables which represent numerical values. Regression algorithms learn to map the input features to a continuous numerical value.
1. Logistic regression 2. Support vector machine 3. Random forest 4. Decision tree 5. K-Nearest Neighbours (KNN) 6. Naive Bayes	**Algorithms**	1. Linear regression 2. Polynomial regression 3. Ridge regression 4. Lasso regression 5. Decision tree 6. Random forest
1. Classifying emails as spam or not spam - Spam detection and email filtering 2. Predicting whether a patient has a high risk of heart disease 3. Yes or No 4. True or False 5. Male or Female	**Example**	1. Predicting the price of a house based on its size, location and amenities 2. Forecasting the sales of a product

- **Advantages of Supervised Machine Learning –**
  1. **High accuracy –** due to training on labelled data
  2. **Process of decision making –** often interpretable
  3. **Often used in pre trained models –** saves times and resources when developing new models from scratch

- **Disadvantages of Supervised Machine Learning –**
  1. **Limitations** in knowing patterns
  2. May struggle with **unseen or unexpected patterns** that are not present in the training data
  3. **Time-consuming and costly –** relies on labelled data only
  4. May lead to **poor generalisations** based on new data

2. Unsupervised Machine Learning –

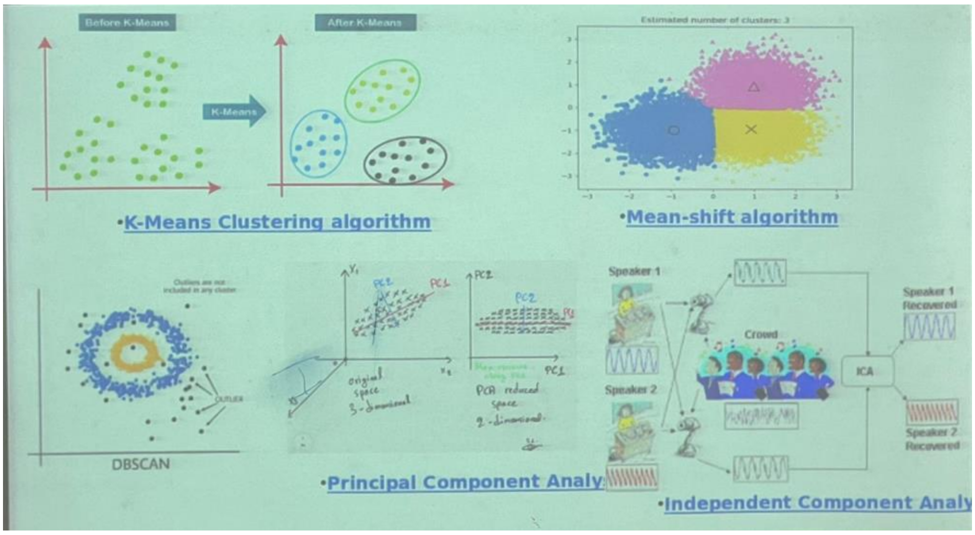
- **Def** – type of machine learning technique in which an algorithm discovers patterns and relationships using unlabelled data
- **Difference between supervised learning and unsupervised learning** – UML doesn't involve providing the algorithm with labelled target outputs
- **Primary goal of UML** – to discover hidden patterns / similarities / clusters within the data → which can then be used for various purposes such as –
  a. data exploration
  b. visualisation
  c. dimensionality reduction etc

- **Example 1** –
  Input data set – Images of a fruit filled container (images not known to machine learning model)
  Task of model – to identify the pattern of objects such as colour, shape or differences seen in the input images and categorise them.
  Conclusion – Upon categorization, the machine then predicts the output as it gets tested with a test data set.

2 main categories of Unsupervised learning

Category of Unsupervised Learning	Definition	Algorithms	Definition of Algorithm
Clustering	• Process of grouping data points into clusters based on their similarity. • Useful for identifying patterns and relationships in data without the need for labelled examples.	1. K – Means Clustering Algorithm	• An unsupervised learning algorithm which groups the unlabelled data set into different clusters. • K = number of predefined clusters that need to be created in the process • Eg – K = 2 → means two clusters • K = 3 → means three clusters and so on
		2. Mean – shift Algorithm	• A centroid – based algorithm that helps in various use cases of unsupervised learning. • Works by – shifting data points towards centroids to be the mean of other points in the region. • One of the best algorithms to be used in image processing and computer vision.

		3. DBSCAN Algorithm	• Clusters are dense regions in the data space, separated by regions of the lower density of points.
		4. Principle Component Analysis (PCA)	• Dimensionality reduction method • Often used to reduce the dimensionality of large data sets ↓ by transforming a large set of variables into a smaller one that still contains most of the information in the large set
		5. Independent Component Analysis (ICA)	• Attempts to decompose a multivariate signal into independent non – Gaussian signals
Association rule learning	• Technique for discovering relationships between items in a data set. • Identifies rules that indicate the presence of one item implies the presence of another item with a specific probability.	1. Apriori Algorithm	• An algorithm for frequent item set mining and association rule learning over relational databases. • Proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those items sets appear sufficiently often in the database.
		2. Eclat Algorithm (Equivalence Class Clustering and Bottom – Up Lattice Traversal)	• Depth – first search – based approach to find frequent item sets. • Relies on the concept of an **‘equivalence class’** to reduce the search space efficiently.
		3. FP – growth Algorithm	• A popular method for frequent pattern mining in data mining. • Works by constructing a frequent pattern tree (FP – tree) from the input data set. • FP – tree is a compressed representation of the data set that captures the frequency and association information of the items in the data.

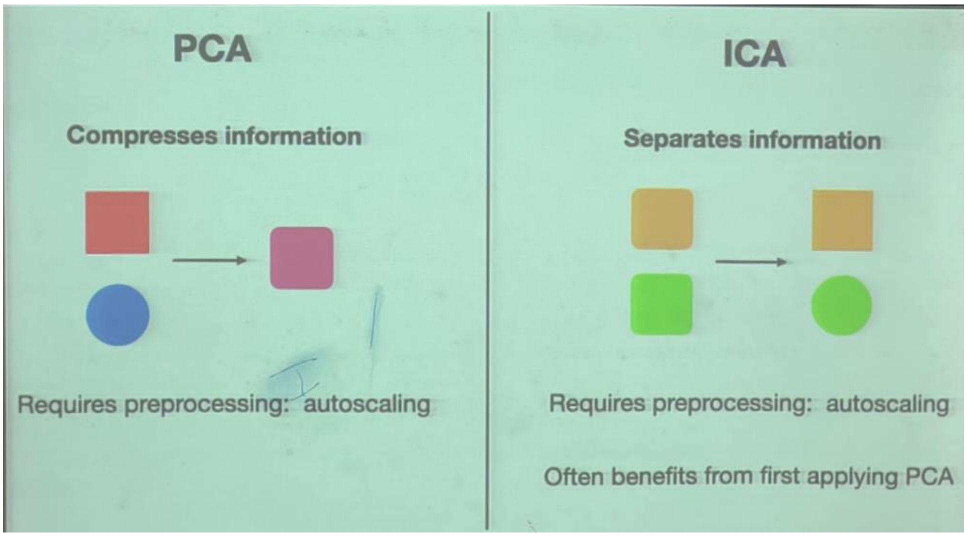


•K-Means Clustering algorithm

•Mean-shift algorithm

•Principal Component Analy

•Independent Component Analy



PCA

Compresses information

Requires preprocessing: autoscaling

ICA

Separates information

Requires preprocessing: autoscaling

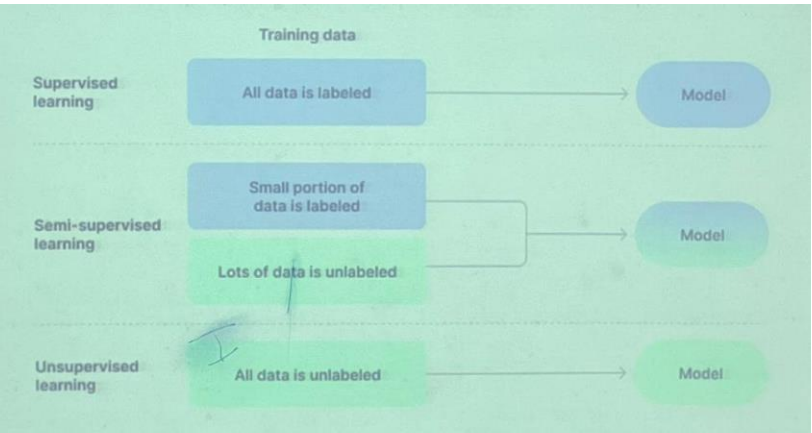
Often benefits from first applying PCA

- **Advantages of UML –**
  1. Helps to discover hidden patterns and various relationships between the data
  2. Used for tasks such as –
     (a) Customer segmentation
     (b) Anomaly detection
     (c) Data exploration
  3. Does not require labelled data → reduces the effort of data labelling

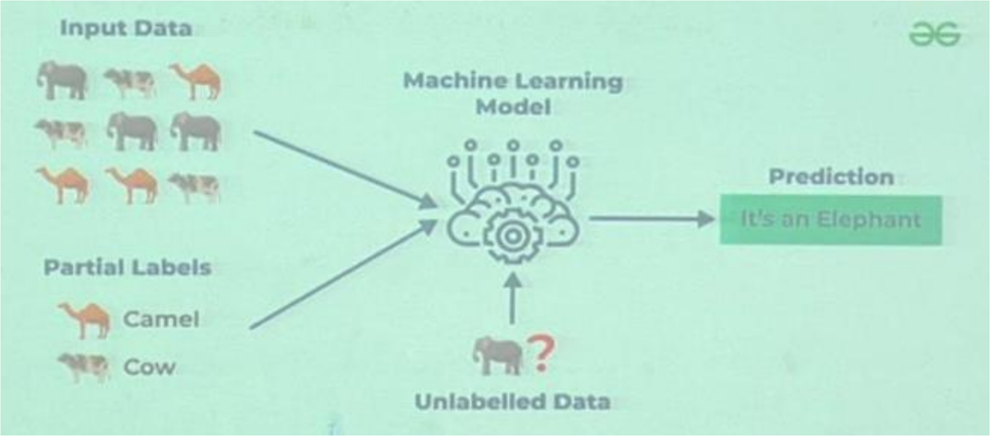
- **Disadvantages of UML –**
  1. Without using labels → may be difficult to predict the quality of the model's output
  2. Cluster Interpretability may not be clear → may not have meaningful interpretations

3. Semi – Supervised Machine Learning –

- **Def –** Machine learning algorithm that works between the supervised and unsupervised learning → uses both labelled and unlabelled data.
- **Particularly useful when –**
  1. Obtaining labelled data is → costly, time – consuming or resource – intensive
  2. The data set is expensive and time – consuming
- SSL chosen when → labelled data requires skills and relevant resources in order to train or learn from it.
- Technique used when dealing with data that is a little bit labelled and the rest large portion of it is unlabelled.
- We can use the unsupervised techniques to predict labels and then feed these labels to supervised methods.



- **Example – Building a language translation model**
  Having labelled translations for every sentence can be resource – intensive
  SSL Allows the models to learn from labelled and unlabelled sentence pairs, making them more accurate.
  This technique has led to significant improvements in the quality of machine translation services.



Types of SSL	Definition
1. Graph – based SSL	- This approach uses a graph to represent the relationships between the data points. ↓ - The graph is then used to propagate labels from the labelled data points to the unlabelled data points.
2. Label propagation	- This approach iteratively propagates labels from the labelled data points to the unlabelled data points, based on the similarities between the data points.
3. Self – training	- This approach trains a machine learning model on the labelled data and then uses the model to predict labels for the unlabelled data. ↓ - The model is then retrained on the labelled data and the predicted labels for the unlabelled data.
4. Co – training	- This approach trains two different machine learning models on different subsets of the unlabelled data. - The 2 models are then used to label each other's predictions. - Derived from the self – training approach and being its improved version, co training is another SSL technique used when only a small portion of label data is available. - Example – when you have two views: one containing image data and the other one audio data → with the co training technique, train the model on audio data and image data separately. Then combine the two models to label the unlabelled data.
5. Generative Adversarial Networks (GANs)	- A type of deep learning algorithm that can be used to generate synthetic data. - GANs used to generate unlabelled data for SSL by training two neural networks: (a) a generator (b) a discriminator

- **Advantages of SSL –**
  1. Leads to better generalization as compared to supervised learning → as it takes both labelled and unlabelled data.
  2. Can be applied to a wide range of data.

- **Disadvantages of SSL –**
  1. SSL Methods can be more complex to implement compared to other approaches.
  2. Still requires some labelled data that might not always be available or easy to obtain
  3. Unlabelled data can impact the model performance accordingly.

4. Reinforcement Learning –

- **Def –** A machine learning training method based on rewarding desired behaviours and punishing undesired ones.
- A reinforcement learning agent (the entity being trained) is able to perceive and interpret its environment, take actions and learn through trial and error → hence also known as '**learning from trials and errors' method.**